Human Genome Project Facts

Contents:

Overview

The Human Genome Project (HGP) is an international research effort to chart and characterize the human genome -- the entire package of genetic instructions for a human being. That entails laying out - in order -- the 3 billion DNA letters (or base pairs) of the full human genetic code.

A great profusion of discoveries about the genetic basis of a long list of diseases already has resulted from the HGP. Initially these discoveries related to relatively rare conditions, but increasingly the same powerful approaches are uncovering hereditary factors in diabetes and other common illnesses.

These revelations hold promise for transforming medical practice. In the years ahead, it may be possible to learn about individual susceptibilities to common disorders such as cancer and heart disease, allowing the design of programs of effective, individualized preventive medicine focused on lifestyle changes, diet and medical surveillance to keep people healthy.

The same discoveries ushered in by the HGP will enable scientists to predict who will respond most effectively to a particular drug therapy, and who may suffer a side effect and ought to avoid that particular drug. In addition, these advances will lead to the next generation of designer drugs, targeted to each individual and engineered in a much more precise way than today's drugs.

As a part of the HGP, 16 research institutions in the United States, Great Britain, Germany, France, Japan, and China are currently generating a high quality, accurate sequence of the human

genetic code for scientists everywhere to use as a no cost resource without restrictions. The project is being done in two stages: the "working draft," and the "finished" sequence. *(See "working draft" and "finished" sequence)*

The government agencies funding the HGP in the U. S. are the National Human Genome Research Institute (NHGRI), a part of the National Institutes of Health (NIH), and the Department of Energy (DOE). Most of the HGP sequencing occurs at four laboratories in the U.S. and one in Great Britain. The three U.S. labs funded by the NHGRI are at Baylor College of Medicine in Houston, Washington University School of Medicine in St. Louis, and Whitehead Institute outside Boston. The DOE sponsors the Joint Genome Institute in Walnut Creek, California. The Wellcome Trust funds the Sanger Centre located outside London.

The HGP international consortium includes two laboratories at the University of Washington in Seattle and labs at Stanford University in Palo Alto, CA; Genome Therapeutics Corp. in Waltham, MA; Genoscope in Evry, France; the RIKEN Institute and Keio University School of Medicine in Japan; and the Max-Planck Institute for Molecular Biology, the Institute of Molecular Biotechnology and the Gesellschaft fuer Biotechnologische in Germany; and the Beijing Human Genome Center in China.

All participants in the HGP have agreed to adhere to specific procedures, including maintaining quality standards and making daily deposits of sequence information into the public databases including GenBank, the European Bioinformatics Institute, and the DNA Database of Japan. Scientists by the thousands, located worldwide, tap into GenBank every day to search for data to advance their medical research.

U.S. HGP has eight research goals:

1. ***Deciphering the human genetic code, or DNA sequence and rapidly providing this data freely and without restrictions to the scientific community and the public***. From 1990 (when the HGP began) until 1996, genetic and physical maps of the human chromosomes and other resources were developed because they were needed to sequence human DNA at relatively low cost and high accuracy. The physical and genetic maps provide landmarks that help scientists navigate the 3 billion pairs of bases, or DNA letters, on the human chromosomes. These maps also have helped scientists hunting for genes even before the "working draft" sequence became available in the public database GenBank. *(See "working draft")* ***In 1996, the HGP sponsored a pilot-sequencing program*** to develop and test methods for large-scale or major DNA sequencing. These efforts were successful, and the ***full-scale effort to sequence the human genome was launched in March 1999***. *(See "sequencing")*
2. Developing ***efficient technology*** to sequence human DNA.
3. Identifying the variations in the human genetic code that underlie disease susceptibility, particularly the most common variations that are called ***SNPs*** (single nucleotide polymorphisms).

4. Interpreting the function of DNA sequence on a genomic scale (***functional genomics***) - determining how individual genes and groups of genes work together in health and disease.
5. Deciphering and analyzing the genetic code of ***model organisms*** such as yeast, roundworm, fruitfly and mouse. The availability of DNA sequence from such organisms expedites scientists' efforts to identify the roles of human genes.
6. Examining the ***ethical, legal and social implications (ELSI)*** of genome research, identifying barriers to the integration of the results of the HGP into health care, and proposing and implementing solutions as appropriate.
7. Developing ***bioinformatic tools and computational strategies*** for the collection, analysis, annotation and storage of the ever-increasing amounts of DNA mapping and sequencing and gene expression data.
8. Training scientists for genomic research and analysis.

Sequencing:

Sequencing means determining the exact order of the base pairs in a segment of DNA. Human chromosomes range in size from about 30,000,000 to 300,000,000 base pairs. There are four different chemical bases, also called nucleotides. They are adenine, thymine, guanine and cytosine, which are abbreviated "A," "T," "G" and "C". The two strands or threads that compose the double helix structure of DNA are essentially strings of these bases. The "As" on one strand always pair with "Ts" on the other strand. And, the "Gs" always pair with "Cs." A base pair is "A" and "T," or "C" and "G." Because the bases exist as pairs, and the identity of one of the bases in the pair determines the other member of the pair, scientists do not have to sequence both bases of the pair.

Whose DNA?:

This is intentionally not known to protect the volunteers who provided DNA samples for sequencing. The sequence is derived from the DNA of several volunteers. To ensure that the identities of the volunteers cannot be revealed, a careful process was developed to recruit the volunteers and to collect and maintain the blood samples that were the source of the DNA.

The volunteers responded to local public advertisements near the laboratories where the DNA "libraries" were prepared. Candidates were recruited from a diverse population. The volunteers provided blood samples after being extensively counseled and giving their informed consent. About five to 10 times as many volunteers donated blood as were eventually used, so that not even the volunteers would know whether their sample was used. All labels were removed before the actual samples to be used were chosen. The samples, without any information about the individual volunteers, were then transferred to a laboratory for construction of DNA clone libraries.

BAC-based sequencing:

The primary method used by the HGP to decipher the human genetic code. BAC is an acronym for "bacterial artificial chromosome." Human DNA is fragmented into pieces that are relatively large but still manageable in size (between 150,000 and 200,000 base pairs). The fragments are cloned in bacteria, which store and replicate the human DNA so that it can be prepared in quantities large enough for sequencing. If carefully chosen to minimize overlap, it takes about 20,000 different BAC clones to contain the 3 billion pairs of bases of the human genome. A collection of BAC clones containing the entire human genome is called a "BAC library."

In the BAC-based method, each BAC clone is "mapped" so that the HGP knows where the DNA in BAC clones comes from on a human chromosome. Thus, the scientists know precisely the location of the DNA letters that are sequenced from each clone and its spatial relation to the human DNA in other BAC clones.

For sequencing, each BAC clone is cut into still smaller fragments, which are about 2,000 bases in length. These pieces are called "subclones." A "sequencing reaction" is carried out on these subclones. The products of the sequencing reaction are then loaded into the sequencing machine (sequencer).

The sequencer produces about 500-800 base pairs of "A," "T," "C," and "G"s in length from each sequencing reaction. A computer can assemble these short sequences into contiguous stretches of sequence representing the human DNA in the BAC clone (*see "depth of coverage").*

Not all areas of the chromosomes can be cloned with current technology. Those areas result in gaps in the sequence. It is expected that those will be uncommon.

Pilot sequencing projects: A set of projects that were initiated in 1996 by the HGP to test the feasibility of deciphering human DNA rapidly, efficiently, and on a large-scale. These projects lasted three years, and their success demonstrated that sequencing the human genome was feasible.

Large-scale sequencing: Cost-effective DNA sequencing conducted on an industrial scale at a rate that is sufficient to generate the sequence of a genome as large as that of the human in a short time. Large-scale sequencing also is characterized by "high throughput". *(see "depth of coverage")*

"Working draft" sequence: intermediate stage in the generation of a high quality, "finished" sequence. "Working draft" sequence is defined as an average of 4X coverage *(see "depth of coverage")*

In early 1999, experiments assessing the usefulness of DNA sequence at various depths of coverage revealed that 4X "working draft" sequence coverage from BAC was extremely useful to biomedical researchers. Thus, HGP consortium leaders decided to pursue a strategy that

would generate "working draft" coverage first, so that scientists would have data for their research as soon as possible. Even though it is not "finished," the "working draft" sequence is being used by scientists throughout the world to speed up their gene-discovery research activities. *(see "finished" sequence)*

"Working draft" sequence that is 4-5X in depth can be assembled into units (called "sequence contigs") that are 10,000 to 12,000 bases in length on average. Although the sequence itself still contains gaps and uncertainties, the sequence contigs are long enough for gene discovery and other biomedical research (see gene discovery), the "working draft" sequence data are deposited into GenBank and other genome sequence databases where access is unrestricted. As a result, scientists are able to use the data now rather than having to wait for the sequence to be "finished".

Although the "draft" version is very useful, the "finished" (the absolute best that humans and computers can accomplish) version will be even more useful and so, after June 2000, the HGP's priority will be to convert the "working draft" to "finished" sequence.

sequencing rate of HGP: 1,000 bases of raw sequence per second, or 12,000 bases of "working draft" per minute. Twenty years ago, deciphering that many bases would have required one year or more. Three years ago, when pilot sequencing projects to evaluate feasibility of human DNA sequencing were initiated, deciphering 12,000 bases required 20 minutes.

"depth of coverage": this refers to the number of times the DNA in a chromosome region is sequenced. A depth of 1 (1X) means that, on average, a particular base pair has been sampled once; a depth of 4 (4X) means that, on average, a particular base has been sequenced four times over. Sequencing the same region many times decreases the possibility of errors in the DNA sequence. Current sequencing instruments can decipher about 500 to 800 bases at a time in a single sequencing "run." The results from these individual "runs" have to be assembled into contiguous stretches of sequence to reconstruct the sequence of a chromosomal region. To build up an accurate assembly from the 500-800 base pair stretches of DNA sequence that emerge from the machines, HGP scientists repeatedly sequence random fragments from each chromosome. *(See BAC-based sequencing and assembly.)* Repeated sequencing allows assembly of much larger regions of DNA because the random individual "runs" overlap with each other, creating areas of commonality that allow the scientists to align the short chunks of DNA sequence into long contiguous sequences.

The **average "depth of coverage" of the HGP's sequence** across the human genome in GenBank is 6 to 7 X. This includes "finished" sequence (9 to 10 X); deep shotgun (8 to 10X) and "working draft" (4 to 5 X).

In addition, the **average "depth of coverage" of clones** across the human genome is estimated to be 32 X. *(See BAC based sequencing)*

scientific publication of "working draft": Analysis has begun on the "working draft" sequence. The results will be submitted for peer review and publication in a prominent scientific journal or journals. Peer review provides final assurance that the research that was conducted is of high quality, interest and utility, and meets scientific standards.

"*Finished" sequence:*

Although "working draft" sequence allows for the recognition of most genes, the higher accuracy and completeness of the "finished" sequence makes it a gold standard. Ready during or before 2003, it will be 99.99 percent accurate, and the only gaps that will remain will be from regions that are impossible to clone or sequence by current methods. All such gaps will be identified and annotated in the database; the gap size will be determined; and all of the efforts made to close them will be recorded.

Thus far, about 20% of the human genome, including two entire chromosomes (21 and 22), has been "finished" to these high standards.

"Finishing" involves performing additional sequencing to increase the "depth of coverage" from the 4-5X of the "working draft" sequence to a total coverage of 9-10X, sometimes referred to as "full shotgun". In addition, finishing requires selecting and sequencing some additional clones from the 10% of the genome that was not included in the "working draft." The data are then inspected by a skilled scientist, aided by computer, to identify any detectable errors or incomplete or missing information. Additional data are specifically collected to bring those regions up to standard. Based on the HGP's experience, the scientists in the HGP consortium likely will conclude the "finishing" within the next three years and possibly sooner.

Assembly:

Putting together the sequenced stretches of DNA into a continuous string of "A", "T", "C" and "Gs" on the chromosome. The sequences are assembled from the individual random 500-800 base runs by computers that compare each of the hundreds of thousands of individual runs and find those that overlap. To be statistically significant, the overlaps must be relatively long, on the order of at least 50 to 100 bases. Thus, if two 500 base runs have an overlap (a shared sequence) of 100 bases, they can be assembled into a longer sequence of 900 bases. By doing this kind of assembly over and over, very long sequences can be built.

The "working draft" is assembled in a two-step fashion. Extensive "fingerprinting" data on each clone allows neighboring BAC clones to be identified. Using the map information about each clone's location, the many BAC clones derived from a chromosome can then be assembled together into a layout of the entire chromosome. *(See BAC-based sequencing and "depth of coverage".)*

HGP scientists constructed the first comprehensive layout of the human genome in mid-May 2000. The layout shows the chromosomal positions and the detailed relationships among the more than 20,000 large clones, which together cover an estimated 97 percent of the euchromatic portion of the genome. It also spotlights the segments remaining to be covered. The clones in the layout also have immense value beyond their immediate role as an aid in sequencing. They provide a permanent resource for human genetics research because they can be used for direct biological studies of gene function.

The euchromatic portion excludes certain regions consisting of long stretches of highly repetitive DNA that encode little genetic information and that are not recovered in the vector systems used by the HGP.

Gene Discovery:

Using computers, scientists can analyze DNA sequence data and recognize the regions with the genes, which encode protein-determining information. Because each portion of the "working draft sequence" is derived from a clone of known location, the locations of the genes that are identified are pinpointed to high resolution in the sequence. The location of a gene that causes a particular disease, or determines an interesting trait, can be compared with the location of the genes that have been identified by computer in the "working draft" sequence in order to determine the exact identity of the disease gene.

"Working draft" sequence already has proven valuable to identifying genes for breast cancer susceptibility (BRCA2); hereditary deafness (Pendred syndrome); several hereditary skeletal disorders; hemorrhagic stroke; focal segmental glomerulosclerosis, a puzzling kidney disorder that can lead to end-stage kidney failure; hereditary epilepsy; and one type of diabetes.

In addition, in clinical trials is a drug for leukemia that was developed based on information in the sequence. Preliminary reports about the drug are very positive.

"Working draft" sequence also has been used to identify over 150,000 sites of variation in the sequence - called single nucleotide polymorphisms -- which are powerful tools for studies of human disease and evolution. A bounty of scientific papers over the next several years will be based on research conducted with "working draft" sequence.

GENBANK (www.ncbi.nlm/gov)

1. A public database of nucleotide sequence operated by the National Center for Biotechnology Information at the National Institutes of Health. The information in GenBank is available to all without restriction. GenBank is a member of an international consortium of nucleotide sequence databases, along with the European Bioinformatics Institute and the DNA Database of Japan. These databases exchange information daily, so that all sequence information is available from any database.

HGP scientists deposit sequence assemblies of 1000 to 2000 bases into one of these databases daily. In the last month, the “working draft” sequence has been flowing into the public databases at a rate of 10,000 DNA letters per minute.

Participation in the HGP does not confer special privileges regarding using or seeking patent applications on data.

Each day about 75,000 searches of GenBank occur.

In addition to human sequence data, GenBank contains DNA sequence of such model organisms such as yeast, worm, fruitfly and many bacteria and other microbes.